Published online 24 September 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D605–D609***
doi:10.1093/nar/gkr796

BμG@Sbase—a microbial gene expression and comparative genomic database

Adam A. Witney[1,*], Denise E. Waldron[1], Lucy A. Brooks[1], Richard H. Tyler[1], Michael Withers[1], Neil G. Stoker[2], Brendan W. Wren[3], Philip D. Butcher[1] and Jason Hinds[1]

[1]Bacterial Microarray Group, Centre for Infection & Immunity, Division of Clinical Sciences, St George's, University of London, Cranmer Terrace, London, SW17 0RE, UK, [2]Department of Pathology and Infectious Diseases, Royal Veterinary College, Royal College Street, London NW1 0TU, UK and [3]Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK

Received August 15, 2011; Revised and Accepted September 9, 2011

ABSTRACT

The reducing cost of high-throughput functional genomic technologies is creating a deluge of high volume, complex data, placing the burden on bioinformatics resources and tool development. The Bacterial Microarray Group at St George's (BμG@S) has been at the forefront of bacterial microarray design and analysis for over a decade and while serving as a hub of a global network of microbial research groups has developed BμG@Sbase, a microbial gene expression and comparative genomic database. BμG@Sbase (http://bugs.sgul.ac.uk/bugsbase/) is a web-browsable, expertly curated, MIAME-compliant database that stores comprehensive experimental annotation and multiple raw and analysed data formats. Consistent annotation is enabled through a structured set of web forms, which guide the user through the process following a set of best practices and controlled vocabulary. The database currently contains 86 expertly curated publicly available data sets (with a further 124 not yet published) and full annotation information for 59 bacterial microarray designs. The data can be browsed and queried using an explorer-like interface; integrating intuitive tree diagrams to present complex experimental details clearly and concisely. Furthermore the modular design of the database will provide a robust platform for integrating other data types beyond microarrays into a more Systems analysis based future.

INTRODUCTION

Despite advances in health treatment, pathogenic bacteria represent one of the most important threats to human health worldwide. Attention has focused on a range of problems, including the alarming spread of antibiotic resistance, microbial contamination of food, the threat of bioterrorism, the global resurgence of tuberculosis, and other emerging and re-emerging infections triggered by lifestyle, political, economical and ecological changes. Research in bacteria has fundamentally changed in the last few years from a piecemeal approach of characterizing individual determinants to a global analysis of host–pathogen interactions. This status has been fuelled by high-throughput technologies, such as microarrays and more recently next generation sequencing. However, these vital efforts are producing unmanageable amounts of data and so it is imperative that resources are available to enable effective mining of not only the new data generated but also the data that is already available. Following the successful example of the public sequence databases, the European Bioinformatics Institute (EBI) and the National Center for Biotechnology (NCBI) have both developed large scale public repositories (ArrayExpress and GEO respectively) to manage these data; however these resources cannot provide focused support to every research community. Thus there is a need for domain specific database resources that can engage at a more personal level. Indeed, Parkhill *et al*. (1) have recommended a three-tier model for the organization of genomic information resources, in which the Tier 1 database is located in individual research laboratories, collecting and analysing locally generated data; the Tier 3 database is the large public repository; and between the two tiers, providing

*To whom correspondence should be addressed. Tel: 02087250698; Fax: 02086720234; Email: awitney@sgul.ac.uk

bidirectional data flow and aggregating domain specific knowledge and tools, are the mid-level Tier 2 database resources. This structured model should allow Tier 2 resources to provide added value to information flowing to Tier 3 and present more domain specific tools to the users of Tier 1.

The biological context of any experiment is essential if the data is to be of further use outside its originating laboratory. In order to communicate this context, standards have been developed to provide a checklist of information that needs to be provided with any data set when published or submitted to a public repository. The Minimum Information about a Microarray Experiment (MIAME) (2) standard has become the requirement for most journals publishing microarray data; this has been followed by other standards, all now described under the umbrella of the Minimum Information for Biological and Biomedical Investigations (MIBBI) project (3).

The bacterial microarray group at St George's (BμG@S), funded by the Wellcome Trust for the last decade, is a community-wide functional genomics resource, that has been designing and producing whole genome microarrays for 16 important bacterial pathogens. In parallel to designing, producing and testing over 30 000 microarrays the BμG@S group was tasked to develop a system to provide domain specific database and analysis tools to the worldwide microbial community that it serves (4). BμG@Sbase is a web-browsable, expertly curated, MIAME-compliant database that acts as a microbial domain specific resource filling the space between the bench research scientist and ArrayExpress/GEO. It enables bench scientists to deposit directly all information relating to experimental design and data output and provides public searchable access to published data.

IMPLEMENTATION

BμG@Sbase has been implemented using all free, open source software. The database backend uses PostgreSQL (http://www.postgresql.org) and the web interface is designed using a combination of PHP (http://www.php.net), Perl (http://www.perl.org), the Javascript libraries jQuery (http://jquery.com) and Overlib (http://www.bosrup.com/web/overlib/), and is served using the Apache web server (http://httpd.apache.org). All systems are hosted using the Linux operating system.

DATABASE DESIGN

The database design of BμG@Sbase was influenced by the Microarray and Gene Expression Object Model (MAGE-OM) (http://www.mged.org/Workgroups/MAGE/mage-om.html) and so is able to capture all aspects of the experimental process, including experimental design, sample treatments, data analysis/transformation and protocol information. A recursive approach to sample/treatment and data/transformation object associations allows the capture of either simple or highly complex experimental procedures. The advantage of this approach is that the exact biological context of the experiment can be captured and displayed, such that the reader can clearly identify all the steps performed if trying to reproduce the experiment and/or analysis. Each process step is also associated with protocol descriptions and parameters, further providing descriptive power to the database content; indeed the database is essentially used as an electronic lab book. Additionally all objects in the database are annotated using a microbial subset of the MGED Ontology (MO) (5), ensuring data is consistently described within the database and can be compared to external data resources that are also using MO.

The database has a mature and extensively tested security model that allows individual research groups to complete fine-grained control over the privacy and sharing of their own data. Read/write permission to any data object can be granted to any user or group, thus allowing users to share data with collaborators before publication. Published data, however, is publicly available and searchable.

Protocols are used throughout the database to annotate all procedural steps; they can be associated with any number of parameters or hardware/software. Protocols can be created on a per lab basis with a set of parameter fields (designated as required or not), which can be populated when the protocol is added to an experimental process.

DATABASE CONTENT

The database currently (September 2011) contains data from 210 experimental investigations of 16 different bacterial pathogens (including the genera: Campylobacter, Clostridium, Francisella, Haemophilus, Helicobacter, Listeria, Mycobacterium, Neisseria, Staphylococcus, Streptococcus, Yersinia); 86 of these are published and therefore publicly available (Table 1). The data has been generated by microbial research groups worldwide with a range of interests (the BμG@S project forms the hub of a collaborative network of over 80 research groups and 250 individual scientists). The research scientists individually enter the data into BμG@Sbase with assistance from the BμG@S curators ensuring uniformity and high quality experimental data curation to internationally recognized standards.

Data entry is provided through a set of web-enabled forms, which guide the user through the process of providing MIAME-compliant data. The forms also allow for efficient annotation of data using microbial specific MO terms. Several convenience tools have also been developed to assist the user in this process; sample

Table 1. BμG@Sbase data statistics (September 2011)

Number of Experiments (Public)	86
Number of Experiments (Total)	210
Number of hybridizations (Public)	2474
Number of hybridizations (Total)	4452
Number of Species with data	35
Number of Bacterial Array designs	58

(Note that in BμG@Sbase, an experiment is equivalent to the datasets contained within a publication).

description information can be edited and uploaded in Microsoft Excel spreadsheet format, and a custom standalone software application, the TIFFreader, is available to be run by the users on their personal computers. The TIFFreader parses both the microarray raw data text and TIFF image files in order to firstly read scanning parameters and other software information, and secondly to check that the correct data files (text and TIFF) are being associated. The user then uploads the resulting Zip file of data to BμG@Sbase. Both these tools ease the process of uploading large amounts of information and reduce the likelihood of errors being made. A submission tool has been developed that will export either a new array design description or a full experimental data set in the Microarray and Gene Expression Markup Language (MAGE-ML) (http://www.mged.org/Workgroups/MAGE/mage-ml.html). Data in this format is submitted to ArrayExpress upon publication of the experiment or pre-publication at the request of the author. Currently 80 published data sets have been submitted to ArrayExpress.

The database contains complete array design information describing all 58 BμG@S microarray designs; these include multi-species/strain pan-genome, species/strain-specific, capsular typing and whole genome tiling designs representing 16 bacterial species. Information regarding the design and manufacture of BμG@S PCR product custom arrays is stored, including PCR primer sequences, reaction conditions, product size and quality, and array print run dates and glass slide types (manufacturer, coatings etc). In addition, BμG@S designs arrays for the Agilent Inkjet *in-situ* Synthesized (IJISS) system thus all oligonucleotide and gene mapping information is stored; the database is used to automatically generate fully annotated mapping files that are available when the arrays are supplied.

Presentation of genome context is enabled using an integrated local instance of the genome browser GBrowse (http://gmod.org/wiki/GBrowse). Currently the database contains information for over 260 bacterial genomes and 119 bacterial plasmids in addition to genomic mapping details of all BμG@S array designs and cross-genome gene BLAST similarities.

DATA EXPLORER INTERFACE

BμG@Sbase uses a web interface to provide maximum flexibility to users without concerns regarding local software installations or upgrades. The interface has been designed to act as an experiment explorer, guiding the user through the various steps of the experimental process (in BμG@Sbase an experiment is equivalent to a publication). From the overall experiment design, including title, abstract, authorship and experimental factors, the user can link through a schematic summary tree (Figure 1A) to view the sample and treatment descriptions (Figure 1B), the protocols used, an array design graphical explorer (Figure 1C), the raw and analysed data files associated or a schematic data analysis tree (Figure 1D).

Sample treatment steps in the context of biological and technical replication can become complicated, but it is essential for the proper understanding of the experimental design that these are presented clearly; thus interactive sample trees are generated (Figure 1B). These are collapsible and expandable at all levels of the tree as well as on a

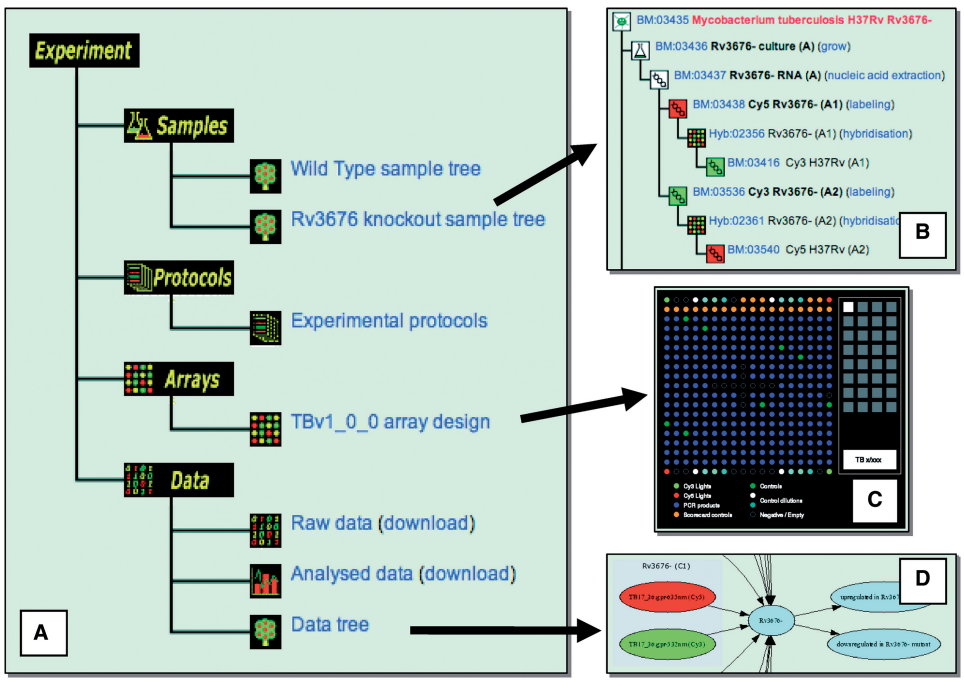

Figure 1. BμG@Sbase stores all experimental data and displays an overview schematic on each experiment page (**A**), sample descriptions are displayed as interactive trees (**B**), an interactive viewer provides access to array design information (**C**), raw and analysed data is linked through a network of transformations (**D**).

node-by-node basis and clearly show each step of the sample treatments in association with a link to its specific protocol and parameters that were used. A general convention is used within the database that replicates are suffixed with a tag (e.g. 'A1') where the letter corresponds to the biological replicate and the number corresponds to the technical replicate.

The graphical Array browser (Figure 1C) allows the user to explore the layout of the array in more detail, searching for further information about the reporters printed; including the PCR primers used for template amplification on the PCR product arrays or the selected sequences for the newer oligonucleotide or IJISS arrays. Supporting layout files for various software applications (e.g. ImaGene, BlueFuse) are also available for download.

The original raw data generated by the feature extraction software can be downloaded as a Zip file (including a description HTML file) from the main experiment summary tree (Figure 1A). Alternatively the data files can be browsed through a list view and individual details accessed, including the feature extraction software versions and settings, quality control metrics and links to any analyses processes in which the file was included. An experiment-wide quality control summary report is available (Figure 2) which provides a series of metrics to identify any problems that may exist in the raw data files. These include checks to make sure that the printed slide has been scanned in the correct orientation, the Cy3 or Cy5 dyes have not been swapped, the overall range of intensity of the elements on the array (a 'traffic light'-like system allows quick visualization of all data files in an experiment) and to identify blank rows in a data file (this provides evidence of a corrupted or edited raw data file). Analysed data can also be downloaded as a Zip file, or browsed in list view; data is stored in column form and/or its original raw format (e.g. images, Microsoft Powerpoint files). These are linked from the raw data through a series of successive transformation steps, each annotated with a protocol. Typically the normalized data for each array is stored followed by the output of any number of filtering steps. The exact analysis process can be visualized on a schematic data tree, also linked from the experimental summary tree (Figure 1D). Analysed data in Newick format tree files can be viewed using the integrated tree viewer applet, PhyloWidget (6). PhyloWidget enables display of phylogenetic trees in several forms (unrooted, circular etc) as well as tree editing, external tree file loading and publication-quality tree image export.

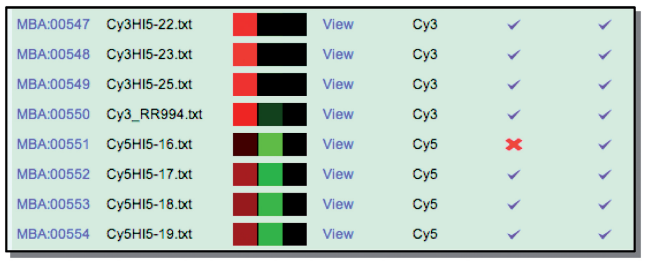

Figure 2. Each experiment has a Quality Control summary report. From left to right: (i) Data file ID, (ii) data file name, (iii) 'traffic light'-like QC display (iv) 'view' link to display details of the calculations, (v) test for correct channel identification, (vi) test for correct array orientation when scanned, (vii) test for significant signal intensity (foreground signal intensity is greater than three standard deviations above background intensity). The 'traffic light'-like system allows quick visualization of all data files; red means a higher percentage of spots are of lower intensity (<1000), green shows a better intensity range (1000 < intensity < 10 000), the third box suggests if there is signal saturation in the data set.

DATA MINING

The data can be queried in several ways through a centralized search page (Figure 3A). Experimental/publication details can be queried to identify studies of interest; these include text based searches of the publication title, abstract or author list as well as MO controlled vocabulary searches of the experimental design or factor types. The experiment can also be queried by the sample descriptions contained within; species, strain, genotype etc can all be queried to identify studies of interest.

Array designs can be identified by organism and design version, and subsequently visualized using the array viewer as previously mentioned; array annotation can then be searched through the 'Spot Search' function. Alternatively a user-supplied sequence can be searched using BLAST against array reporters (PCR products or oligonucleotides).

Analysed data can be searched by first identifying a gene of interest, either through the systematic name, gene name or some associated attribute (e.g. GO term or PFAM domain identifier). A gene summary page (Figure 3B) presents annotation data with a link to a genome browser view (Figure 3D) as well as links to external database resources. An embedded query form provides a search of database wide gene lists (e.g. up or down-regulated lists) or normalized data (filtering on ratio, log ratio or *P*-value if required); matching data sets are returned grouped by experiment (Figure 3C). If gene lists are searched, a comparison function enables a further search of all these matching lists and identifies other genes, which are also present; ordering by occurrence can identify genes which may show similar patterns to the gene of interest.

FUTURE DIRECTIONS

While originally developed for microarray data, the modular infrastructure developed for recording biological sample information, processing raw data and integrating complex analysis tools behind an intuitive web-interface, is ideal for effective expansion to additional genomics technology platforms and other Systems biology approaches. Thus development to integrate other forms of microbial functional genomic data, e.g. RNA-Seq, ChIP-Seq is currently underway.

An updated interface is in development, harnessing the power of Web 2.0 drag-and-drop technologies, which will create a more dynamic user experience and provide access to a suite of analysis tools that are also in development. The analysis tools will also be used to perform a

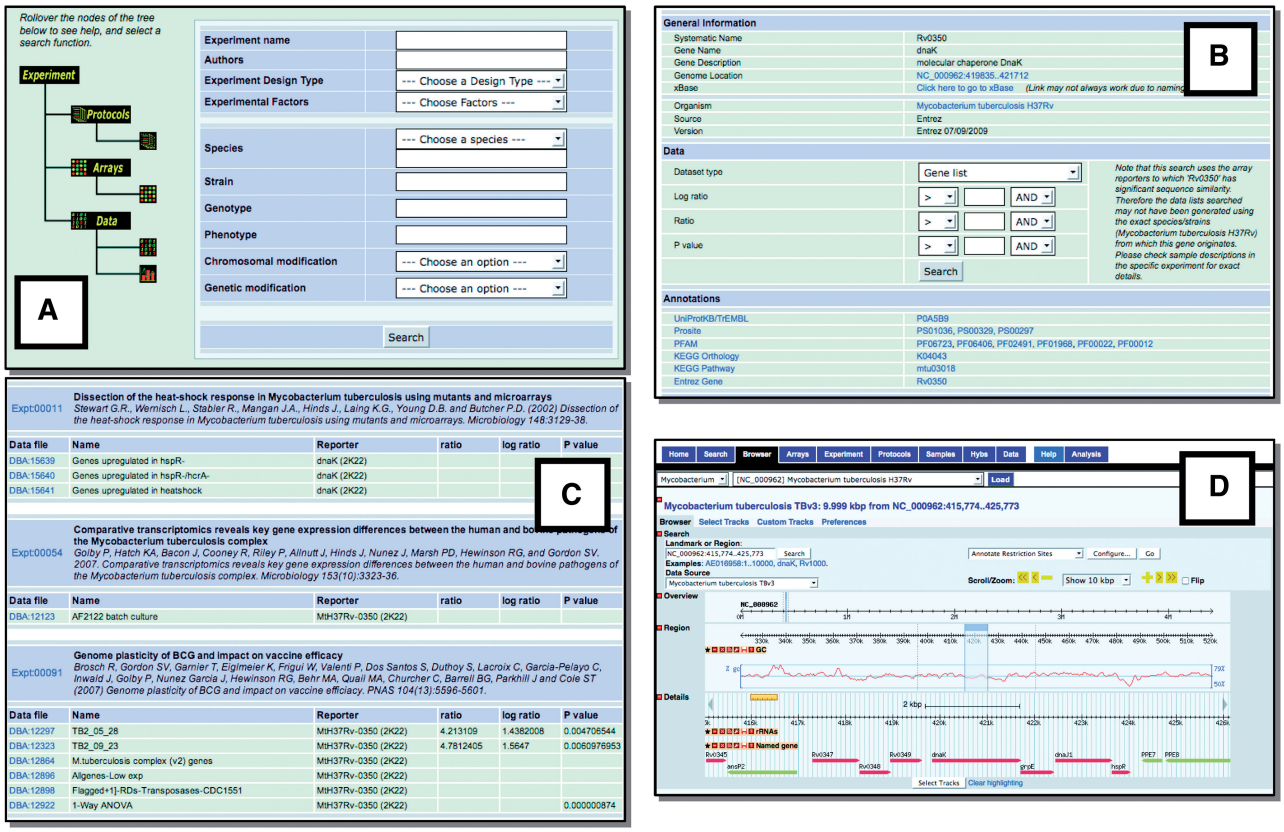

Figure 3. Representative example screenshots. The database can be queried from a single search page (**A**). Gene-centric queries first identify a gene of interest with associated annotation information (**B**) which can then be used to search analysed data and the resulting gene lists or normalized files grouped by experiment (**C**). The gene can also be visualized in genome context using a local instance of GBrowse (**D**).

re-analysis of the data sets to provide a uniform platform for inter-experimental comparison.

In order to promote interactivity between BμG@Sbase and other databases, a RESTful Application Programming Interface (API) is also in development. This will provide better communication with other databases and also allow developers of other resources to be able to query BμG@Sbase directly from within their own database applications and tools. Using this API, MAGE-TAB (7) formatted data will be exported, further improving database interactivity, a key attribute as we move towards a systems level analysis of biological data.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the valuable feedback provided by BμG@Sbase users, which helps in both the identification of bugs and the development of new features.

FUNDING

The Wellcome Trust under a Functional Genomics Resources Initiative and a Biomedical Resources grant (grant numbers 062511, 080039, 086547). Funding for open access charges: The Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Parkhill,J., Birney,E. and Kersey,P. (2010) Genomic information infrastructure after the deluge. *Genome Biol.*, **11**, 402–406.
2. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al.* (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
3. Taylor,C.F., Field,D., Sansone,S.A., Aerts,J., Apweiler,R., Ashburner,M., Ball,C.A., Binz,P.A., Bogue,M., Booth,T. *et al.* (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.*, **26**, 889–896.
4. Witney,A.A. and Hinds,J. (2002) BμG@Sbase – a microarray database and analysis tool. *Comparative Funct. Genomics*, **3**, 369–371.
5. Whetzel,P.L., Parkinson,H., Causton,H.C., Fan,L., Fostel,J., Fragoso,G., Game,L., Heiskanen,M., Morrison,N., Rocca-Serra,P. *et al.* (2006) The MGED ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics*, **22**, 866–873.
6. Jordan,G.E. and Piel,W.H. (2008) PhyloWidget: web-based visualizations for the tree of life. *Bioinformatics*, **24**, 1641–1642.
7. Rayner,T.F., Rocca-Serra,P., Spellman,P.T., Causton,H.C., Farne,A., Holloway,E., Irizarry,R.A., Liu,J., Maier,D.S., Miller,M. *et al.* (2006) A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, **7**, 489.